

GENE NETWORK RECONSTRUCTION USING GLOBAL-LOCAL SHRINKAGE PRIORS*

BY GWENAËL G.R. LEDAY[†], MATHISCA C.M. DE GUNST[‡], GINO B.
KPOGBEZAN[§], AAD W. VAN DER VAART[§], WESSEL N. VAN
WIERINGEN^{‡,¶} AND MARK A. VAN DE WIEL^{‡,¶}

*MRC Biostatistics Unit[†], Vrije Universiteit Amsterdam[‡],
Leiden University[§] and VU University Medical Center[¶]*

Reconstructing a gene network from high-throughput molecular data is often a challenging task, as the number of parameters to estimate easily is much larger than the sample size. A conventional remedy is to regularize or penalize the model likelihood. In network models, this is often done *locally* in the neighbourhood of each node or gene. However, estimation of the many regularization parameters is often difficult and can result in large statistical uncertainties. In this paper we propose to combine local regularization with *global* shrinkage of the regularization parameters to borrow strength between genes and improve inference. We employ a simple Bayesian model with non-sparse, conjugate priors to facilitate the use of fast variational approximations to posteriors. We discuss empirical Bayes estimation of hyper-parameters of the priors, and propose a novel approach to rank-based posterior thresholding. Using extensive model- and data-based simulations, we demonstrate that the proposed inference strategy outperforms popular (sparse) methods, yields more stable edges, and is more reproducible.

1. Introduction. Gaussian Graphical Models (GGMs) are a popular tool in genomics to describe functional dependencies between biological units of interest, such as genes or proteins. These models provide means to apprehend the complexity of molecular processes using high-throughput experimental data, and shed light on key regulatory genes or proteins that may be interesting for further follow-up studies. Among the many approaches that have been advanced, simultaneous-equation models (SEMs), which express each gene or protein expression profile as a function of other ones, have been found particularly valuable owing to their flexibility and simplicity. Notably, SEMs facilitate *local* regularization, where for each gene the

*This work was supported by the Center for Medical Systems Biology (CMSB), established by the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NGI/NWO), and the European Union Grant EpiRadBio, nr. FP7-269553.

Keywords and phrases: Undirected gene network, Bayesian inference, Shrinkage, Variational approximation, Empirical Bayes

set of parameters that model its dependence on the other genes is penalized separately and possibly to a different amount. However this comes at the price of having many regularization parameters, which may be difficult to tune. Motivated by works in the field of differential expression analysis, in this paper we combine local regularization with *global* shrinkage of the regularizing parameters to stabilize and improve estimation. Adopting a Bayesian approach, we demonstrate, using extensive model- and data-based simulations, that such global shrinkage may substantially improve statistical inference.

GGMs characterize the dependence structure between molecular variables using partial correlations. It is well known that two coordinates Y_i and Y_j of a multivariate normal random vector $Y = (Y_1, \dots, Y_p)^T$ are conditionally independent given the set of all other coordinates if and only if the partial correlation $\text{corr}(Y_i, Y_j | Y_{\mathcal{J} \setminus \{i,j\}})$ is zero, where $\mathcal{J} = \{1, \dots, p\}$. Furthermore, if $Y \sim \mathcal{N}_p(0, \Omega^{-1})$ with positive-definite *precision matrix* $\Omega = (\omega_{ij})$, then these partial correlations can be expressed as $\text{corr}(Y_i, Y_j | Y_{\mathcal{J} \setminus \{i,j\}}) = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}}$, for $i \neq j$. Thus the conditional dependence structure is fully coded in the precision matrix, and a network structure may be defined by discriminating the zero and non-zero entries of the precision matrix. It is convenient to represent this structure by an undirected graph $\mathcal{G} = \{\mathcal{J}, \mathcal{E}\}$, with the nodes \mathcal{J} corresponding to the variables, and the edge set \mathcal{E} consisting of all $\{i, j\}$ such that $\omega_{ij} \neq 0$.

Most modern inference techniques for GGMs focus on estimating Ω or this underlying graph. For brevity we only discuss the most popular methods, which will also be used as benchmarks in our simulations.

Penalized likelihood estimation amounts to maximizing $\ell(\Omega) = \log |\Omega| - \text{tr}(S\Omega) - \lambda J(\Omega)$, where S is the sample covariance estimate, J a penalty function, and λ a scalar tuning parameter. The penalty J may serve two purposes: (1) to ensure identifiability and improve the quality of estimation; (2) to discriminate zero from non-zero entries in Ω . The ℓ_1 -norm (or versions thereof) is a popular choice (Friedman, Hastie and Tibshirani, 2008), because it simultaneously achieves (1) and (2). Alternatively, a ridge-type penalty (van Wieringen and Peeters, 2014; Warton, 2008; Ledoit and Wolf, 2004) may be used in combination with a thresholding procedure (Schäfer and Strimmer, 2005). Appropriate tuning of the penalty through the parameter λ is crucial for good performance. Various solutions, usually based on resampling or cross-validation, have been proposed (Gao et al., 2012; Lian, 2011; Meinshausen and Bühlmann, 2010; Foygel and Drton, 2010; Giraud, 2008; Yuan and Lin, 2007).

Simultaneous-equation modelling estimates Ω by regressing each molecu-

lar variable Y_j against all others. The coefficients $\beta_{j,k}$ in the equations

$$(1.1) \quad Y_j = \sum_{k \in \mathcal{J} \setminus j} Y_k \beta_{j,k} + \epsilon_j, \quad j \in \mathcal{J},$$

where $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ is independent of $(Y_k : k \neq j)$, can be shown to be given by $\beta_{j,k} = -\omega_{jj}^{-1} \omega_{jk}$. Also $\sigma_j^2 = \omega_{jj}^{-1}$. Consequently, identifying the nonzero entries of Ω can be recast as a variable selection problem in p Gaussian regression models. This approach to graphical modeling was popularized by [Meinshausen and Bühlmann \(2006\)](#). They dealt with high-dimensionality by adding an ℓ_1 -penalty to each regression problem, but other penalties are also used ([Krämer, Schafer and Boulesteix, 2009](#)). Because the model (1.1) misses the symmetry $\omega_{ij} = \omega_{ji}$ in Ω , estimation may lack efficiency. This may be overcome by working directly on partial correlations, as shown by [Peng, Zhou and Zhu \(2009\)](#). Alternatively, [Meinshausen and Bühlmann \(2006\)](#) proposed a *post-symmetrization* step with an ‘AND’ rule: edge $(i, j) \in \mathcal{E}$ if $\beta_{i,j} \neq 0$ and $\beta_{j,i} \neq 0$. Despite the symmetry issue, network reconstruction using (1.1) performs well and is widely used in practice.

Simultaneous-equation models are quite flexible. Experimental or biological covariates can easily be accounted for in the regression, and extensions to non-Gaussian data were suggested by [Chen, Witten and Shojaie \(2015\)](#); [Allen and Liu \(2013\)](#); [Yang et al. \(2012\)](#); [Ravikumar, Wainwright and Lafferty \(2010\)](#). Also SEMs arise naturally from the differential equations of a general dynamical system model of gene regulation ([Oates and Mukherjee, 2012](#)).

In this paper we develop a Bayesian approach to Gaussian graphical modeling using SEMs. Our contribution is three-fold: (1) we employ (1.1) in combination with (non-sparse) priors that induce both *local* and *global* shrinkage and provide evidence that global shrinkage may substantially improve inference; (2) we present a new approach to posterior thresholding using a concept similar to the local false discovery rate ([Efron, 2010](#)) and show that non-sparse priors coupled with a posteriori edge selection are a simple and attractive alternative to sparse priors; and (3) we provide a computationally attractive software tool called **ShrinkNet** (available at <http://github.com/gleday/ShrinkNet>), which is based on a coherent and complete estimation procedure that does not rely on resampling or cross-validation schemes to tune parameter(s).

The paper is organized as follows. Section 2 presents the Bayesian SEM, the variational approximation to posteriors and a novel posterior thresholding procedure to reconstruct the network. In this section we also describe estimation of the global shrinkage prior and discuss the important role of

the proposed empirical Bayes procedure, along with its connection to existing literature. In Sections 3 and 4 we compare the performance of the new method with state-of-the-art sparse and non-sparse approaches, using both model- and data-based simulations. Notably in Section 4 we employ two mRNA expression data sets from The Cancer Genome Atlas (TCGA) and a random-splitting strategy to compare the reproducibility and stability of the various methods.

2. Methods. In this section we introduce the Bayesian SEM with global and local shrinkage priors along with a variational approximation of the resulting posterior distribution(s). Next we present empirical Bayes estimation of prior hyper-parameters. We conclude with a selection procedure for inferring the edge set \mathcal{E} .

2.1. The Bayesian SEM. Consider mRNA expression data on p genes from n sample tissues. Denote by \mathbf{y}_j the $n \times 1$ vector of mRNA expression (\log_2) values for gene $j \in \mathcal{J} = \{1, \dots, p\}$. The Bayesian SEM is defined by equation (1.1) together with a hierarchical specification of prior distributions:

$$\begin{aligned}
 \mathbf{y}_j &= \sum_{k \in \mathcal{J} \setminus j} \mathbf{y}_k \beta_{jk} + \boldsymbol{\epsilon}_j, & j = 1, \dots, p \\
 \boldsymbol{\epsilon}_j &\sim \mathcal{N}_n(0, \sigma_j^2 \mathbf{I}_n), \\
 \beta_{jk} &\sim \mathcal{N}(0, \sigma_j^2 \tau_j^2), \\
 \tau_j^{-2} &\sim \mathcal{G}(a, b), \\
 \sigma_j^{-2} &\sim \mathcal{G}(c, d).
 \end{aligned}
 \tag{2.1}$$

Here (as usual) every line is understood to be conditional on the lines below it and variables within a line are assumed independent, as are variables referring to different genes j . Furthermore, $\mathcal{G}(s, r)$ denotes a gamma distribution with shape and rate parameters s and r , and \mathbf{I}_n is the $n \times n$ identity matrix. Throughout the paper the hyper-parameters c and d are fixed to small values, e.g. 0.001, in contrast to a and b , which we will estimate (see Section 2.3). Although c and d could also be estimated, we prefer a non-informative prior for the parameters σ_j , as there seems no reason to connect the error variances across the equations.

The regression parameters β_{jk} are endowed with gene-specific, Gaussian priors for *local* shrinkage. A small value of the prior variance τ_j^2 encourages the posterior distributions of the β_{jk} (including their expectations $\mathbb{E}(\beta_{jk} | \mathbf{y}_j)$) to be shrunk towards zero. The stabilizing effect of this ridge-type shrinkage has been observed to be useful for ranking regression parameters as a

first step in variable selection (Bondell and Reich, 2012). In Section 2.4 we show how similarly the marginal posterior distributions of the β_{jk} can be used for rank-based edge selection in a GGM. The prior variances of the β_{jk} are also defined proportional to the error variances σ_j^2 to bring the variances τ_j^2 , and the induced shrinkage, on a comparable scale (Park and Casella, 2008).

The equations for different genes j are connected through the gamma priors placed on the precisions τ_j^{-2} and the error variances σ_j^2 , for $j \in \mathcal{J}$. The prior on the error variances has no structural role, and, as mentioned, we prefer a fixed non-informative prior. In contrast, the $\mathcal{G}(a, b)$ -prior on the precisions τ_j^{-2} induces *global* shrinkage by borrowing strength across the regression equations. The *exchangeability* of the precisions expressed through this prior acknowledges the fact that the equations for the different genes are similar in a broad sense, which is plausible given that they share many common elements. When informative (i.e. small or moderate value of a/b^2), this prior shrinks the posterior distributions of τ_j^{-2} towards the prior mean a/b , which stabilizes estimation. This type of shrinkage is different from the shrinkage of the regression coefficients β_{jk} , which through their centered priors are always shrunk to zero. Of course, the “informed” shrinkage of the precisions τ_j^{-2} will be beneficial only if the hyper parameters a and b are chosen appropriately. We propose to set their values based on the data, using an empirical Bayes approach, discussed in Section 2.3.

The conjugacy of the Gaussian and gamma priors in model (2.1) confers the method a computational advantage over complex sparse priors. Fast approximations to the posteriors are readily available (Rajagopalan and Broemeling, 1983; Rue, Martino and Chopin, 2009; Ormerod and Wand, 2010), whereas sparse, non-conjugate priors often require MCMC. The Gaussian priors allow to reparameterize the problem employing an SVD decomposition of the design matrix (West, 2003), and back-transform the posteriors to the original space (at least in our setting with approximately Gaussian posteriors; see Section 2.2), which is computationally advantageous when $p > n$.

A disadvantage of these priors is that they do not have an intrinsic variable selection property, whence the posterior does not automatically recover the graph structure. We solve this by a separate procedure for variable selection, which essentially consists of thresholding the scaled posterior means of the regression coefficients β_{jk} . In Section 2.4 we present an approach based on Bayes factors and a local false discovery rate.

2.2. *Variational approximation to posteriors.* Because intractable integrals make it difficult to obtain the exact posterior distribution of the parameters, we use a variational approximation. Variational inference is a fast deterministic alternative to MCMC methods, and consists of computing a best approximation to the posterior distribution from a prescribed family of distributions. In our situation it provides an analytic expression for a lower bound on the log-marginal likelihood, which is useful for monitoring convergence of the algorithm and to assess model fit (Section 2.3).

For given hyper-parameters (a, b) and with the variables \mathbf{y}_k in the right side of (2.1) considered fixed covariates, the prior and posterior distributions factorize (i.e. are independent) across the genes j . For simplicity of notation we shall omit the index j from τ_j^{-2} , σ_j^{-2} , \mathbf{y}_j and β_j in the remainder of this section. Hence the formulas for $\boldsymbol{\lambda} := (\beta, \tau^{-2}, \sigma^{-2})$ below apply to the joint posterior distribution of $(\beta_j, \tau_j^{-2}, \sigma_j^{-2})$, for (any) given $j \in \mathcal{J}$.

We shall seek a variational approximation to the posterior distribution of $\boldsymbol{\lambda}$ within the class of all distributions with independent marginals over β , τ^{-2} and σ^{-2} , where we measure the discrepancy by the Kullback-Leibler (KL) divergence. Thus letting $p(\boldsymbol{\lambda}|\mathbf{y})$ denote the posterior density in model (2.1), we seek to find a density $q(\boldsymbol{\lambda})$ of the form

$$(2.2) \quad q(\boldsymbol{\lambda}) = q_1(\beta)q_2(\tau^{-2})q_3(\sigma^{-2}),$$

for some marginal densities q_1, q_2, q_3 , that minimizes the Kullback-Leibler divergence

$$(2.3) \quad \begin{aligned} \text{KL}(q||p) &= \int q(\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\lambda})}{p(\boldsymbol{\lambda}|\mathbf{y})} d\boldsymbol{\lambda} \\ &= \mathbb{E}_q \log q(\boldsymbol{\lambda}) - \mathbb{E}_q \log p(\boldsymbol{\lambda}, \mathbf{y}) + \log p(\mathbf{y}), \end{aligned}$$

over all densities q of product form. Here $p(\mathbf{y})$ denotes the marginal density of the observation in model (2.1). Because the Kullback-Leibler divergence is nonnegative we have that

$$(2.4) \quad \mathbb{E}_q \log p(\boldsymbol{\lambda}, \mathbf{y}) - \mathbb{E}_q \log q(\boldsymbol{\lambda}) \leq \log p(\mathbf{y}).$$

Furthermore, minimization of the Kullback-Leibler divergence is equivalent to maximization of the left side of this inequality. Thus we may think of the procedure as maximizing a lower bound on the log marginal likelihood.

The solution q^* of this maximization problem, with the marginal densities q_1, q_2, q_3 left completely free, can be seen to be given by densities q_1^*, q_2^*, q_3^* satisfying (see (Blei and Jordan, 2006; Ormerod and Wand, 2010))

$$(2.5) \quad q_m^*(\boldsymbol{\lambda}_m) \propto \exp \left\{ \mathbb{E} \prod_{m' \neq m} q_{m'} \log p(\boldsymbol{\lambda}, \mathbf{y}) \right\}, \quad m = 1, 2, 3.$$

In the context of our model this yields $q^*(\boldsymbol{\lambda}) = q_1^*(\boldsymbol{\beta})q_2^*(\tau^{-2})q_3^*(\sigma^{-2})$, with the marginal densities (see Section 1 of Supplementary Material (SM)) given by standard distributions,

$$(2.6) \quad \begin{aligned} q_1^*(\boldsymbol{\beta}) &= {}^d \mathcal{N}_{p-1}(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*) \\ q_2^*(\tau^{-2}) &= {}^d \mathcal{G}(a^*, b^*), \\ q_3^*(\sigma^{-2}) &= {}^d \mathcal{G}(c^*, d^*), \end{aligned}$$

where the parameters on the right side satisfy

$$\begin{aligned} \boldsymbol{\beta}^* &= (\mathbf{X}^T \mathbf{X} + \mathbb{E}_{q_2^*}[\tau^{-2}] \mathbf{I}_{p-1})^{-1} \mathbf{X}^T \mathbf{y} \\ \boldsymbol{\Sigma}^* &= [\mathbb{E}_{q_3^*}[\sigma^{-2}] (\mathbf{X}^T \mathbf{X} + \mathbb{E}_{q_2^*}[\tau^{-2}] \mathbf{I}_{p-1})]^{-1} \\ a^* &= a + \frac{p-1}{2}, \\ b^* &= b + \frac{1}{2} \mathbb{E}_{q_3^*}[\sigma^{-2}] \mathbb{E}_{q_1^*}[\boldsymbol{\beta}^T \boldsymbol{\beta}], \\ c^* &= c + \frac{n+p-1}{2}, \\ d^* &= d + \frac{1}{2} \mathbb{E}_{q_1^*}[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] + \frac{1}{2} \mathbb{E}_{q_2^*}[\tau^{-2}] \mathbb{E}_{q_1^*}[\boldsymbol{\beta}^T \boldsymbol{\beta}]. \end{aligned}$$

Here \mathbf{X} represents the n by $p-1$ fixed design matrix of (2.1). For the j^{th} equation in (2.1) this is equal to $\mathbf{y}_{-j} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{j-1}^T, \mathbf{y}_{j+1}^T, \dots, \mathbf{y}_p^T)^T$.

Furthermore, the variational lower bound on the log-marginal likelihood $\log p(\mathbf{y})$ (the left side of (2.4)) evaluated at $q = q^*$ simplifies to:

$$(2.7) \quad \begin{aligned} \mathcal{L} &= -\frac{n}{2} \log(2\pi) + \frac{1}{2} \log |\boldsymbol{\Sigma}^*| + \frac{1}{2} (p-1) + a \log b - \log \Gamma(a) - \\ &\quad a^* \log b^* + \log \Gamma(a^*) + c \log d - \log \Gamma(c) - c^* \log d^* + \\ &\quad \log \Gamma(c^*) + \frac{1}{2} \mathbb{E}_{q_3^*}[\sigma^{-2}] \mathbb{E}_{q_2^*}[\tau^{-2}] \mathbb{E}_{q_1^*}[\boldsymbol{\beta}^T \boldsymbol{\beta}]. \end{aligned}$$

See SM Section 1 for the details.

The equations (2.6) express the optimal densities q_1^* , q_2^* and q_3^* (or equivalently the parameters in the right side of (2.6)) in terms of each other. This motivates a coordinate ascent algorithm (Blei and Jordan, 2006; Ormerod and Wand, 2010) (depicted in Algorithm 1), which proceeds by updating the parameters in turn, replacing the variational densities on the right hand sides of the equations by their current estimates, at every iteration.

Upon convergence the marginal posteriors $p(\boldsymbol{\beta}|\mathbf{y})$, $p(\tau^{-2}|\mathbf{y})$ and $p(\sigma^{-2}|\mathbf{y})$ are approximated by $q_1^*(\boldsymbol{\beta})$, $q_2^*(\tau^{-2})$ and $q_3^*(\sigma^{-2})$. Although the algorithm

needs to be repeated for each regression equation in (2.1), the overall computational cost of the procedure is low.

In high-dimensional settings with $n < p$ it is preferable to use an SVD decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{F}\mathbf{V}^T$ of the regression matrix, and reparameterize the model relative to the $n \times n$ matrix \mathbf{F} , i.e. in terms of the principal components of $\mathbf{X}^T\mathbf{X}$ (West, 2003). Then we compute the variational (Gaussian) approximation $q_1^*(\boldsymbol{\theta})$ to the posterior density of the parameter $\boldsymbol{\theta}$ in this reduced model, and obtain a variational approximation to the marginal posterior density $p(\boldsymbol{\beta}|\mathbf{y})$ of the original parameter by (linearly) transforming back. For the purpose of edge selection (see Section 2.4) we are only interested in the posterior expectation $\mathbb{E}_p[\boldsymbol{\beta}|\mathbf{y}]$ and posterior variance $\mathbb{V}_p[\boldsymbol{\beta}|\mathbf{y}]$ of $\boldsymbol{\beta}$ (which are approximated by $\mathbb{E}_{q_1^*}[\boldsymbol{\beta}]$ and $\mathbb{V}_{q_1^*}[\boldsymbol{\beta}]$), and hence only the diagonal of the high-dimensional posterior covariance matrix needs to be determined.

Algorithm 1 Variational algorithm for local shrinkage

```

1: Initialize:
2:  $b = d = b^{*(0)} = d^{*(0)} = 0.001$ ,  $\xi = 10^{-3}$ ,  $M = 1000$  and  $t = 1$ 
3: while  $|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}| \geq \xi$  and  $2 \leq t \leq M$  do
4:   update  $\Sigma^{*(t)} \leftarrow \left[ \mathbb{E}_{q_3^{*(t-1)}}(\sigma^{-2}) \left( \mathbf{X}^T\mathbf{X} + \mathbb{E}_{q_2^{*(t-1)}}(\tau^{-2})\mathbf{I}_{p'} \right) \right]^{-1}$ 
5:   update  $\boldsymbol{\beta}^{*(t)} \leftarrow \mathbb{E}_{q_3^{*(t-1)}}(\sigma^{-2})\Sigma^{*(t)}\mathbf{X}^T\mathbf{y}$ 
6:   update  $d^{*(t)} \leftarrow d + \frac{1}{2} \left[ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{*(t)})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{*(t)}) + \text{tr}\{\mathbf{X}^T\mathbf{X}\Sigma^{*(t)}\} \right] +$ 
       $\frac{1}{2}\mathbb{E}_{q_2^{*(t-1)}}(\tau^{-2}) \left[ \boldsymbol{\beta}^{*(t)T}\boldsymbol{\beta}^{*(t)} + \text{tr}\{\Sigma^{*(t)}\} \right]$ 
7:   update  $b^{*(t)} \leftarrow b + \frac{1}{2}\mathbb{E}_{q_3^{*(t-1)}}(\sigma^{-2}) \left[ \boldsymbol{\beta}^{*(t)T}\boldsymbol{\beta}^{*(t)} + \text{tr}\{\Sigma^{*(t)}\} \right]$ 
8:   update  $\mathcal{L}^{(t)}$ 
9:    $t \leftarrow t + 1$ 
10: end while

```

2.3. Empirical Bayes and prior calibration. In the preceding discussion we have treated the vector of hyper-parameters $\boldsymbol{\alpha} = (a, b)$ as fixed. We now turn to its estimation and present a modified variational algorithm in which $\boldsymbol{\alpha}$ is updated along with the other parameters. The new algorithm is akin to an EM algorithm (Braun and McAuliffe, 2010) in which the two steps are, respectively, replaced with a variational E-step, where the lower bound is optimized over the variational parameters via coordinate ascent updates, and a variational M-step, where the lower bound is optimized over $\boldsymbol{\alpha}$ with the variational parameters held fixed.

We now use the SEM for all genes together, and write the variational approximation for the posterior density of the parameters for the j th gene as q^j . (For each j this is given by a triple of three marginal densities.) The target is to maximize the sum over the genes of the lower bounds on the log-marginal likelihood, i.e. the sum over j of the left side of (2.4), which can be written as

$$(2.8) \quad \sum_{j=1}^p \mathbb{E}_{q^j} \log p(\mathbf{y}_j | \boldsymbol{\lambda}_j) + \sum_{j=1}^p \mathbb{E}_{q^j} \log \frac{p_{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_j)}{q^j(\boldsymbol{\lambda}_j)} \leq \sum_{j=1}^p \log p_{\boldsymbol{\alpha}}(\mathbf{y}_j).$$

Maximization of the left side with respect to the densities q^j for a fixed hyper-parameter $\boldsymbol{\alpha}$ would lead to the variational estimates q^{j*} given by (2.6). However, rather than iterating (2.6) until convergence, we now alternate between ascending in q and in $\boldsymbol{\alpha}$. For the variational estimates q^j fixed at their current iterates, optimizing the left-hand side of (2.8) relative to the parameter $\boldsymbol{\alpha}$ amounts to maximizing, with the current iterate q^{j*} replacing q^j ,

$$(2.9) \quad \sum_{j=1}^p \mathbb{E}_{q^{j*}} \log p_{\boldsymbol{\alpha}}(\tau_j^{-2}) = \sum_{j=1}^p (a \log b - \log \Gamma(a) + (a-1) \mathbb{E}_{q^{j*}} \log \tau_j^{-2} - b \mathbb{E}_{q^{j*}} \tau_j^{-2}).$$

The exact solution to this problem can be found using a fixed-point iteration method, as in (Valpola and Honkela, 2006). Alternatively, the following approximate solution arises by analytical maximization after replacing the digamma function $\psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$ by the approximation $\log(x) - 0.5x^{-1}$:

$$(2.10) \quad \begin{cases} \hat{a} = \frac{1}{2} \left[\log \left(\sum_{j=1}^p \mathbb{E}_{q^{j*}} \tau_j^{-2} \right) - p^{-1} \sum_{j=1}^p \mathbb{E}_{q^{j*}} \log \tau_j^{-2} - \log p \right]^{-1} \\ \hat{b} = \hat{a} \cdot p \cdot \left[\sum_{j=1}^p \mathbb{E}_{q^{j*}} \tau_j^{-2} \right]^{-1} \end{cases}$$

Algorithm 2 outlines how the updates of the hyper-parameters are folded into the variational algorithm. At iteration t the hyper-parameters $a^{(t)}$ and $b^{(t)}$ are computed according to (2.10) with the expectations $\mathbb{E}_{q^{j*}} \tau_j^{-2}$ and $\mathbb{E}_{q^{j*}} \log \tau_j^{-2}$ computed under the current estimates q^{j*} . Next the variational parameters defining the densities q^{j*} are updated according to (2.6) using the values $a^{(t)}$ and $b^{(t)}$ for a and b . Figure 1(a) illustrates the convergence of the algorithm and shows that the lower bound on the sum of log-marginal likelihoods increases at each step of the algorithm (red line). Although this

is not true for the lower bounds of each regression equation in the SEM, this does demonstrate that the estimation procedure yields a well-informed prior that is beneficial overall.

The second summand on the left-hand side of (2.8) is equal to minus $\sum_{j=1}^p \text{KL}(q^{j*} || p_{\alpha})$. This suggests that the procedure will seek to set the hyper parameters α so that the prior density p_{α} of the λ_j on the average most resembles their (approximate) posteriors q^{j*} , based on the different genes. This connects to the recent work of [Van de Wiel et al. \(2013\)](#) on shrinkage priors for differential gene expression analysis, whose empirical Bayes procedure consists in finding α such that $p_{\alpha}(\tau_j^{-2}) \approx n^{-1} \sum_j p_{\alpha}(\tau_j^{-2} | \mathbf{y}_j)$. Figure 1(b) shows that our approach fulfills the same objective. It is natural for the empirical Bayes procedure to have this “averaging of marginal posteriors” property, as it attempts to calibrate the prior according to the data. The role of the global shrinkage prior $\mathcal{G}(a, b)$ is to encourage the posterior distributions of the τ_j^{-2} , for $j \in \mathcal{J}$, to shrink to a common distribution, centered around the (prior) mean a/b .

Algorithm 2 Variational EM algorithm with global-local shrinkage priors

```

1: Initialize:
2:  $a^{(0)} = b^{(0)} = a^{*(0)} = 0.001, \forall j \in \mathcal{J}, b_j^{*(0)} = d_j^{*(0)} = 0.001, \xi = 10^{-3}, M = 1000$  and  $t = 1$ 
3: while  $\max |\mathcal{L}_j^{(t)} - \mathcal{L}_j^{(t-1)}| \geq \xi$  and  $2 \leq t \leq M$  do
    E-step: Update variational parameters:
4:   for  $j = 1$  to  $p$  do
5:     update  $\Sigma_j^{*(t)}, \beta_j^{*(t)}, d_j^{*(t)}, b_j^{*(t)}$  and  $\mathcal{L}_j^{(t)}$  in that order (as in Algorithm 1)
6:     update  $a^{*(t)} \leftarrow a^{(t-1)} + \frac{p-1}{2}$ 
7:   end for
    M-step: Update hyper-parameters:
8:    $a^{(t)} \leftarrow 0.5 \left( p^{-1} \sum_{j=1}^p \left( \log(b_j^{*(t)}) - \psi(a^{*(t)}) \right) - \log p + \log \sum_{j=1}^p \frac{a^{*(t)}}{b_j^{*(t)}} \right)^{-1}$ 
9:    $b^{(t)} \leftarrow a^{(t)} \cdot p \left( \sum_{j=1}^p \frac{a^{*(t)}}{b_j^{*(t)}} \right)^{-1}$ 
10:   $t \leftarrow t + 1$ 
11: end while

```

2.4. Edge selection. In this section we describe a separate procedure for edge selection. This consists of first ranking the edges based on summary statistics from the (marginal) posterior distributions under the model (2.1) obtained in the preceding sections, and next performing forward selection along this ordering. For the latter we use Bayes factors and their relation to a Bayesian version of the local false discovery rate ([Efron, 2010](#), lfdr).

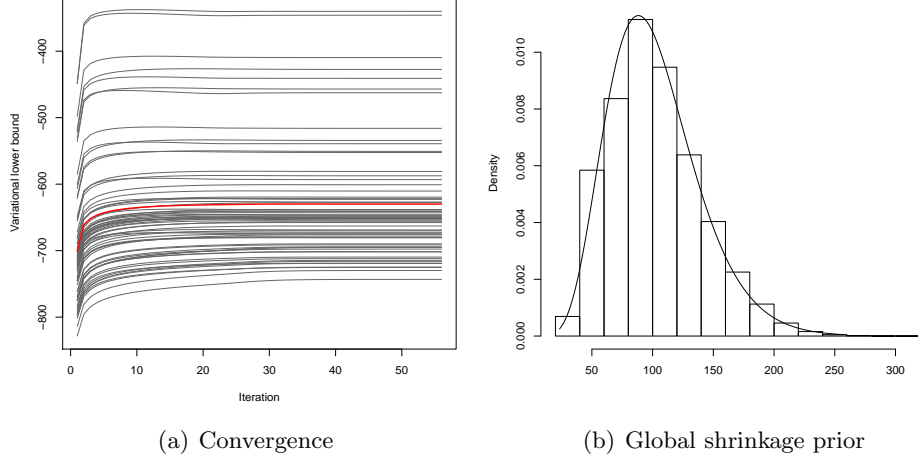


FIG 1. Illustration of (a) the convergence of the variational algorithm and (b) the estimated global shrinkage prior on the breast cancer data set (P53 pathway). Figure (a) displays the variational lower bounds \mathcal{L}_j of each regression equation in the SEM as a function of iterations. The red continuous line represents the average lower bound. Figure (b) displays an empirical mixture of marginal posteriors of τ_j^{-2} obtained by drawing 1000 samples from $q_2^j(\tau_j^{-2}; \mathbf{y}_j)$, $\forall j$. The continuous line represents the density of the estimated global shrinkage prior on τ_j^{-2} , which correspond to $\mathcal{G}(7.404, 0.073)$.

2.4.1. *Edge ordering.* Denote the approximate posterior expectation and variance of $\beta_{j,k}$ obtained in Sections 2.2 and 2.3 for SEM (2.1) by $\mathbb{E}_{q^{j*}}[\beta_{j,k}|\mathbf{y}_j]$ and $\mathbb{V}_{q^{j*}}[\beta_{j,k}|\mathbf{y}_j]$, and define

$$(2.11) \quad \kappa_{j,k} = \frac{|\mathbb{E}_{q^{j*}}[\beta_{j,k}|\mathbf{y}_j]|}{\sqrt{\mathbb{V}_{q^{j*}}[\beta_{j,k}|\mathbf{y}_j]}}, \quad j, k \in \mathcal{J} \text{ with } j \neq k.$$

Next for a given edge (j, k) (between genes j and k) define the quantity $\bar{\kappa}_{j,k} = (\kappa_{j,k} + \kappa_{k,j})/2$, and order the set of $P = p(p-1)/2$ edges according to their associated values $\bar{\kappa}_{j,k}$, from large to small. Let $(j(r), k(r))$ denote the r th edge in this ordering, and abbreviate its associated value to $\bar{\kappa}_r = \bar{\kappa}_{j(r), k(r)}$. This ordering is retained in all of the following. However, we do not necessarily select all edges below a certain threshold, but proceed by forward selection, for $r = 1, \dots, P$.

2.4.2. *Bayes factors.* Selection at stage r (see Section 2.4.5) will be based on Bayes factors $\text{BF}(j(r), k(r))$ and $\text{BF}(k(r), j(r))$ for the two regression parameters $\beta_{j(r), k(r)}$ and $\beta_{k(r), j(r)}$ associated with the r th edge.

Denote by $m_{j(r),k(r),1}$ the model in SEM (2.1) for the response variable $\mathbf{y}_{j(r)}$, with the covariates (or nonzero $\beta_{j(r),k}$) restricted to the edge $(j(r), k(r))$ and any *previously selected* edge (involving node $j(r)$) with rank lower or equal to $r - 1$. Likewise, define $m_{j(r),k(r),0}$, but with the restriction $\beta_{j(r),k(r)} = 0$, which is equivalent to the exclusion of edge $(j(r), k(r))$. The Bayes factor associated with this model is

$$(2.12) \quad \text{BF}(j(r), k(r)) = \frac{p(\mathbf{y}_{j(r)} | m_{j(r),k(r),1})}{p(\mathbf{y}_{j(r)} | m_{j(r),k(r),0})}, \quad r = 1, \dots, P.$$

The Bayes factor $\text{BF}(k(r), j(r))$ is defined analogously from the regression models $m_{k(r),j(r),1}$ and $m_{k(r),j(r),0}$ for response variable $\mathbf{y}_{j(k)}$.

2.4.3. Prior for Bayesian variable selection. The global shrinkage prior for the precision parameters τ_j^{-2} estimated from the data in Section 2.3 is not appropriate for computing the Bayes factors (2.12). Because it has been calibrated (by the variational Bayes method outlined in Algorithm 2) for the network comprised of all edges, it is likely to be located away from zero, which will induce strong regularization on the regression parameters, making it difficult for the Bayes factors to discriminate between the subsequent models (in particular when n is small). A non-informative prior runs into the same problem (perhaps even in a more severe manner).

Motivated by the Zellner-Siow prior (Liang et al., 2008; Zellner and Siow, 1980) we propose to employ instead the “default prior” $\tau_j^{-2} \sim \mathcal{G}(1/2, n/2)$. This concentrates near its prior expectation n^{-1} (i.e. the fixed unit information prior of Kass and Wasserman (1995)), and hence is concentrated near 0 for moderate and large values of n , while less stringent for small n (see illustration in SM Section 4).

2.4.4. Bayesian analogue of lfr. Since both Bayes factors $\text{BF}(j(r), k(r))$ and $\text{BF}(k(r), j(r))$ are informative for the relevance of edge $(j(r), k(r))$, we need to combine these and find a suitable threshold. For that purpose, we link the Bayes factors to the posterior null-probability $P_0(\bar{\kappa}_r) = P(\beta_{j(r),k(r)} = 0, \beta_{k(r),j(r)} = 0 | \mathbf{y})$, where $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_p^T)^T$. The absence of edge $(j(r), k(r))$ is reflected by $\beta_{j(r),k(r)} = \beta_{k(r),j(r)} = 0$, which, in the spirit of forward selection, implies the null models $m_{j(r),k(r),0}$ and $m_{k(r),j(r),0}$. The posterior null-probability is linked to the local false discovery rate (Efron, 2010, lfr). However, as in Van de Wiel et al. (2013), we condition on the data \mathbf{y} rather than on a test statistic. Then, we have

$$(2.13) \quad \begin{aligned} P_0(\bar{\kappa}_r) &= P(\beta_{j(r),k(r)} = 0, \beta_{k(r),j(r)} = 0 | \mathbf{y}) \\ &\leq \min\{P(\beta_{j(r),k(r)} = 0 | \mathbf{y}), P(\beta_{k(r),j(r)} = 0 | \mathbf{y})\}. \end{aligned}$$

Here, the bound is used because the SEM may not provide accurate joint probabilities on regression coefficients from different regression models. Now, assume the prior null probability $P(\beta_{j,k} = 0|\mathbf{y}_{-j}) = p_0, \forall j \in \mathcal{J}$, where $\mathbf{y}_{-j} = (\mathbf{y}_1^T, \dots, \mathbf{y}_{j-1}^T, \mathbf{y}_{j+1}^T, \dots, \mathbf{y}_p^T)^T$. Note that a constant value of p_0 is reasonable, because it simply reflects the prior probability that response \mathbf{y}_j does not respond to covariate \mathbf{y}_k (which is a member of \mathbf{y}_{-j}). Then,

$$\begin{aligned}
 P(\beta_{j,k} = 0|\mathbf{y}) &= P(\beta_{j,k} = 0|\mathbf{y}_j, \mathbf{y}_{-j}) \\
 &= \frac{P(\mathbf{y}_j|\beta_{j,k} = 0, \mathbf{y}_{-j})P(\beta_{j,k} = 0|\mathbf{y}_{-j})}{P(\mathbf{y}_j|\mathbf{y}_{-j})} \\
 (2.14) \quad &= \frac{P(\mathbf{y}_j|m_{j,k,0})p_0}{P(\mathbf{y}_j|m_{j,k,0})p_0 + (1-p_0)P(\mathbf{y}_j|m_{j,k,1})} \\
 &= \frac{p_0}{p_0 + (1-p_0)\text{BF}(j, k)}.
 \end{aligned}$$

Define the max Bayes factor: $\text{BF}(\bar{\kappa}_r) = \max\{\text{BF}(j(r), k(r)), \text{BF}(k(r), j(r))\}$. Then, after substituting (2.14) into (2.13) we have, for threshold $\gamma = (1 - \alpha)p_0/(\alpha(1 - p_0))$,

$$(2.15) \quad \text{BF}(\bar{\kappa}_r) \geq \gamma \iff P_0(\bar{\kappa}_r) \leq \alpha.$$

Equation (2.15) suggests that edges in the graph can be selected using a thresholding rule on the Bayes factors that controls the posterior null-probability. For example, when we have $p_0 = 0.9$, then $\text{BF}(\bar{\kappa}_r) > 81$ implies $P_0(\bar{\kappa}_r) < 0.1$. However, to use this approach an estimate of p_0 is required. We simply propose

$$(2.16) \quad \hat{p}_0 = \frac{1}{2P} \left(\sum_{r=1}^P (I_{\{\text{BF}'(j(r), k(r)) \leq 1\}} + I_{\{\text{BF}'(k(r), j(r)) \leq 1\}}) \right).$$

where $\text{BF}'(j(r), k(r))$ is defined analogously to $\text{BF}(j(r), k(r))$, but *without* forward selection (so all covariates corresponding to edge ranks $\leq r$ are included), because the forward selection procedure requires knowing \hat{p}_0 .

2.4.5. Forward selection procedure. We introduce the following sequential procedure to update the set \mathbf{E} of selected edges and the models $m_{j(r), k(r), 0}$, $m_{j(r), k(r), 1}$, $m_{k(r), j(r), 0}$, $m_{k(r), j(r), 1}$ when increasing r :

1. Initiate α , $r = 1$, $\ell = 0$ and $\mathbf{E}^0 = \emptyset$. Compute γ from α and \hat{p}_0 .
2. Determine the models $m_{j(r), k(r), 0}$ and $m_{k(r), j(r), 0}$ which are the current models for $\mathbf{y}_{j(r)}$ and $\mathbf{y}_{k(r)}$ that correspond to \mathbf{E}^{r-1} . Augment those models with covariates $\mathbf{y}_{k(r)}$ and $\mathbf{y}_{j(r)}$, respectively, and fit these models to obtain $m_{j(r), k(r), 1}$ and $m_{k(r), j(r), 1}$.

3. Calculate the max Bayes factor $\text{BF}(\bar{\kappa}_r)$
4. Only if $\text{BF}(\bar{\kappa}_r) > \gamma$ update $\mathbf{E}^r = \mathbf{E}^{r-1} \cup \{(j(r), k(r))\}$
5. Update $r = r + 1$ and go back to step 2

For the purpose of variable selection we include intercepts in the SEM. Finally, we estimate \mathcal{E} by the last update of \mathbf{E} .

The selection procedure respects the initial ranking of the edges in terms of the order in which they are considered for inclusion in the forward selection. However, the procedure is set up to proceed when a given edge is not selected, because in the light of the current model subsequent edges may (slightly) increase the marginal likelihood. As in practice we observed that the Bayes factor decreases with r (see Supplementary Figure 2), a stopping criterion may be practical if P is large; e.g. stop if r reaches $r_{\max} = (1 - \hat{p}_0)P$, or if $\text{BF}(\bar{\kappa}_r)$ has not exceeded γ for, say, 100 consecutive values of r .

3. Model-based simulation. In this section we investigate the performance of our approach, termed ShrinkNet, in recovering the structure of an undirected network and compare it to popular approaches. We generate $n \in \{25, 50, 100\}$ samples from a multivariate normal distribution with mean vector $\mathbf{0}$ and 100×100 precision matrix $\mathbf{\Omega}$, corresponding to four different graph structures: *band*, *cluster*, *hub* and *random* (Zhao et al., 2012) (see Figure 2 for illustration), every of them sparse, with graph density ranging from 0.017 to 0.096. We generated the inverse covariance matrix $\mathbf{\Omega}$ corresponding to each graph structure from a G-Wishart distribution (Mohammadi and Wit, 2015) with scale matrix equal to the identity and $b = 4$ degrees of freedom. In SM Section 2 we provide statistical summaries on the magnitude of the generated partial correlations.

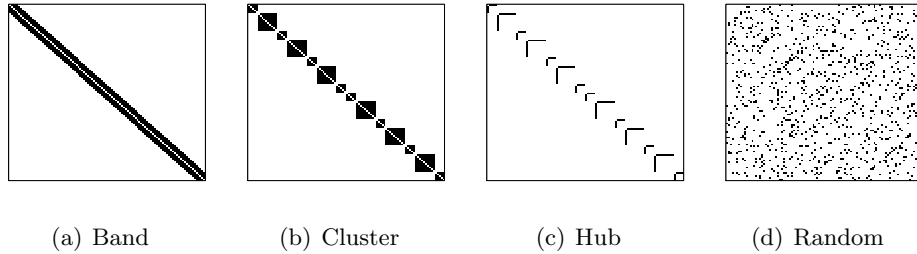


FIG 2. Graph structures considered for the precision matrix $\mathbf{\Omega}$ in our simulation. Black and white dots represent non-zero and zero entries, respectively. Only off-diagonal elements are displayed. For precision matrices with block-diagonal structures (clusters and hubs), block sizes were set to 5 and 10. In (a) the bandwidth is equal to four. The graph density δ is (a) $\delta = 0.079$, (b) $\delta = 0.071$, (c) $\delta = 0.017$ and (d) $\delta = 0.096$.

We compared our approach ShrinkNet to the popular frequentist SEM with the Lasso penalty (SEM_L) (Meinshausen and Bühlmann, 2006), the Graphical Lasso (GL_λ) (Friedman, Hastie and Tibshirani, 2008), and GeneNet (Schaefer, Opgen-Rhein and Strimmer, 2006). The latter combines a non-sparse linear shrinkage estimator with an *a posteriori* edge selection procedure. For the purpose of comparison with ShrinkNet, we also consider the Bayesian SEM (2.1) with the non-informative global shrinkage prior $\mathcal{G}(0.001, 0.001)$, which we subsequently refer to as ‘NoShrink’.

Briefly, graph selection is as follows. For SEM_L and GL_λ we use the EBIC criterion (Foygel and Drton, 2010; Chen and Chen, 2008) for selecting the optimal regularization parameter(s), whereas for GeneNet and ShrinkNet a threshold of 0.1 on the local false discovery rate and the posterior null probability P_0 is employed. In SM Section 3 we provide more details as to how an edge ranking is obtained for each method.

To evaluate the performance of the methods in recovering the graph structures we report partial ROC curves (SM Section 5), which depict the true positive rate (TPR) as a function of the false positive rate (FPR) for $\text{FPR} < 0.2$, and various performance measures on selected graphs. Table 3 below provides a numerical summary: it displays averages and standard deviations of TPRs, FPRs, F-scores and partial AUCs (pAUC) (Dodd and Pepe, 2003) as a function of the method, n and the true graph structure. The $\text{F-score} = 2 \times (\text{precision} \times \text{TPR}) / (\text{precision} + \text{TPR})$ is a popular performance measure, defined as the harmonic mean between the $\text{TPR} = \text{TP} / (\text{TP} + \text{FN})$ (also called *recall*) and the $\text{precision} = \text{TP} / (\text{TP} + \text{FP})$, where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively.

Table 3 shows that ShrinkNet achieves the highest average partial AUCs in all but two situations in which it is very close to the maximum. The results also indicate that NoShrink is often outperformed by GeneNet, and comparable to GL_λ , which suggests that the global shrinkage carried out by ShrinkNet considerably improves edge ranking. SEM_L has the lowest average pAUC in all situations.

The performance of each method in recovering the true graph structure can also be evaluated by the average F-score. According to this metric the best performance is achieved by NoShrink and ShrinkNet in all but two cases. In moderate- ($n = 50$) and high-dimensional cases ($n = 25$), NoShrink and ShrinkNet show a much larger F-score than others. This is particularly pronounced when $n = 25$, in which case GL_λ and GeneNet have an F-score (and TPR) very close to zero. In this context SEM_L is performing better than GL_λ and GeneNet, but worse than NoShrink and ShrinkNet.

	n	Method	TPR	FPR	F-score	pAUC
BAND	100	NoShrink	0.255 (0.014)	0.001 (0.000)	0.404 (0.017)	0.342 (0.007)
		ShrinkNet	0.246 (0.015)	0.001 (0.000)	0.392 (0.020)	0.354 (0.007)
		SEM _L	0.143 (0.013)	0.000 (0.000)	0.249 (0.020)	0.331 (0.006)
		GL _λ	0.144 (0.065)	0.001 (0.001)	0.242 (0.094)	0.334 (0.007)
		GeneNet	0.226 (0.025)	0.001 (0.000)	0.366 (0.032)	0.351 (0.007)
	50	NoShrink	0.150 (0.015)	0.001 (0.001)	0.257 (0.022)	0.292 (0.009)
		ShrinkNet	0.142 (0.013)	0.001 (0.001)	0.245 (0.020)	0.312 (0.010)
		SEM _L	0.059 (0.011)	0.000 (0.000)	0.110 (0.019)	0.273 (0.010)
		GL _λ	0.013 (0.010)	0.000 (0.000)	0.028 (0.018)	0.294 (0.009)
		GeneNet	0.084 (0.020)	0.000 (0.000)	0.154 (0.034)	0.309 (0.009)
	25	NoShrink	0.065 (0.011)	0.002 (0.001)	0.119 (0.019)	0.244 (0.010)
		ShrinkNet	0.066 (0.011)	0.002 (0.001)	0.122 (0.020)	0.259 (0.010)
		SEM _L	0.017 (0.007)	0.000 (0.000)	0.033 (0.014)	0.220 (0.008)
		GL _λ	0.004 (0.004)	0.000 (0.000)	0.011 (0.006)	0.247 (0.009)
		GeneNet	0.000 (0.002)	0.000 (0.000)	0.011 (0.008)	0.260 (0.010)
CLUSTER	100	NoShrink	0.290 (0.015)	0.000 (0.000)	0.449 (0.018)	0.375 (0.008)
		ShrinkNet	0.276 (0.015)	0.000 (0.000)	0.432 (0.018)	0.389 (0.007)
		SEM _L	0.175 (0.014)	0.000 (0.000)	0.297 (0.020)	0.336 (0.007)
		GL _λ	0.402 (0.038)	0.001 (0.001)	0.566 (0.037)	0.357 (0.007)
		GeneNet	0.299 (0.027)	0.000 (0.000)	0.458 (0.031)	0.383 (0.008)
	50	NoShrink	0.174 (0.016)	0.000 (0.000)	0.295 (0.022)	0.326 (0.010)
		ShrinkNet	0.163 (0.015)	0.000 (0.000)	0.279 (0.022)	0.350 (0.010)
		SEM _L	0.078 (0.010)	0.000 (0.000)	0.144 (0.018)	0.280 (0.008)
		GL _λ	0.022 (0.013)	0.000 (0.000)	0.043 (0.025)	0.317 (0.008)
		GeneNet	0.120 (0.026)	0.000 (0.000)	0.213 (0.040)	0.344 (0.009)
	25	NoShrink	0.084 (0.013)	0.001 (0.001)	0.153 (0.021)	0.275 (0.010)
		ShrinkNet	0.086 (0.014)	0.001 (0.001)	0.156 (0.023)	0.295 (0.011)
		SEM _L	0.031 (0.008)	0.000 (0.000)	0.059 (0.015)	0.228 (0.009)
		GL _λ	0.004 (0.005)	0.000 (0.000)	0.013 (0.009)	0.270 (0.008)
		GeneNet	0.004 (0.006)	0.000 (0.000)	0.017 (0.013)	0.293 (0.010)
HUB	100	NoShrink	0.357 (0.031)	0.000 (0.000)	0.520 (0.035)	0.385 (0.014)
		ShrinkNet	0.354 (0.030)	0.000 (0.000)	0.519 (0.034)	0.388 (0.013)
		SEM _L	0.277 (0.027)	0.000 (0.000)	0.431 (0.033)	0.355 (0.013)
		GL _λ	0.459 (0.055)	0.004 (0.001)	0.540 (0.035)	0.384 (0.014)
		GeneNet	0.302 (0.043)	0.001 (0.001)	0.446 (0.047)	0.381 (0.013)
	50	NoShrink	0.241 (0.030)	0.001 (0.000)	0.378 (0.039)	0.346 (0.016)
		ShrinkNet	0.242 (0.029)	0.000 (0.000)	0.383 (0.038)	0.353 (0.015)
		SEM _L	0.149 (0.026)	0.000 (0.000)	0.258 (0.038)	0.310 (0.017)
		GL _λ	0.183 (0.052)	0.001 (0.001)	0.295 (0.067)	0.345 (0.014)
		GeneNet	0.123 (0.034)	0.000 (0.000)	0.216 (0.053)	0.346 (0.015)
	25	NoShrink	0.141 (0.030)	0.001 (0.001)	0.234 (0.044)	0.290 (0.016)
		ShrinkNet	0.147 (0.031)	0.001 (0.001)	0.242 (0.044)	0.307 (0.016)
		SEM _L	0.063 (0.022)	0.000 (0.000)	0.117 (0.039)	0.256 (0.015)
		GL _λ	0.038 (0.028)	0.000 (0.000)	0.080 (0.046)	0.296 (0.017)
		GeneNet	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.302 (0.016)
RANDOM	100	NoShrink	0.230 (0.014)	0.002 (0.001)	0.367 (0.018)	0.317 (0.007)
		ShrinkNet	0.227 (0.013)	0.002 (0.001)	0.365 (0.017)	0.323 (0.006)
		SEM _L	0.132 (0.012)	0.001 (0.000)	0.232 (0.018)	0.309 (0.007)
		GL _λ	0.036 (0.023)	0.000 (0.000)	0.069 (0.042)	0.315 (0.006)
		GeneNet	0.080 (0.017)	0.000 (0.000)	0.148 (0.029)	0.296 (0.008)
	50	NoShrink	0.143 (0.014)	0.004 (0.001)	0.242 (0.021)	0.271 (0.009)
		ShrinkNet	0.136 (0.014)	0.003 (0.001)	0.234 (0.021)	0.283 (0.007)
		SEM _L	0.048 (0.010)	0.000 (0.000)	0.091 (0.018)	0.258 (0.009)
		GL _λ	0.008 (0.005)	0.000 (0.000)	0.018 (0.010)	0.278 (0.007)
		GeneNet	0.023 (0.011)	0.000 (0.000)	0.045 (0.021)	0.262 (0.009)
	25	NoShrink	0.071 (0.011)	0.004 (0.001)	0.127 (0.018)	0.225 (0.009)
		ShrinkNet	0.070 (0.010)	0.004 (0.001)	0.127 (0.017)	0.233 (0.009)
		SEM _L	0.012 (0.005)	0.000 (0.000)	0.025 (0.010)	0.208 (0.009)
		GL _λ	0.002 (0.002)	0.000 (0.000)	0.007 (0.004)	0.234 (0.009)
		GeneNet	0.001 (0.002)	0.000 (0.000)	0.008 (0.005)	0.226 (0.010)

TABLE 1

Average and standard deviations (in parentheses) of performance measures over 100 repetitions. The best performance in average for the F-score and pAUC (partial Area Under the Curve) are boldfaced.

All in all, the simulation study demonstrates that global shrinkage considerably improves edge ranking. For network reconstruction, the small discrepancy between ShrinkNet and NoShrink indicates that the selection procedure of Section 2.4 is relatively robust to edge ranking. The proposed selection procedure is also shown to outperform contenders in the most high-dimensional cases.

4. Data-based simulation. In this section we employ gene expression data from The Cancer Genome Atlas (TCGA) to compare the performance of our approach in reconstructing networks with SEM_L , GL_λ , GeneNet and NoShrink (see previous Section). Data were retrieved from the TCGA cBioPortal using the R package ‘cgdsr’ (Jacobsen, 2013; Cerami et al., 2012). In particular, we focus on the p53 pathway in the Breast cancer data set ($n_{\text{brca}} = 526$), which comprise $p^{\text{p53}} = 67$ genes, and the apoptosis pathway in the Ovarian data set ($n_{\text{ov}} = 537$) that comprises $p^{\text{apopt}} = 79$ genes. Since the true molecular network is not exactly known, we employ a random splitting strategy for the two data sets to assess discoveries.

4.1. Reproducibility. To study reproducibility, we randomly split the data into a small data set where $n_{\text{small}}^{\text{p53}} \in \{134, 67, 34\}$ and $n_{\text{small}}^{\text{apopt}} \in \{158, 79, 40\}$ to achieve low-, moderate- and high-dimensional situations, and a large data set where $n_{\text{large}}^{\text{p53}} \in \{392, 459, 492\}$ and $n_{\text{large}}^{\text{apopt}} \in \{379, 458, 497\}$ (representing the complement). The large data set is then used to validate discoveries made from the small one. As a benchmark for validation we employ the edge set \mathcal{S}_b defined by edges that are simultaneously selected by the different methods based on the large data set. Because the lack of consensus between the different methods may render \mathcal{S}_b too small, we only compare two methods at a time.

To assess performance in recovering \mathcal{S}_b from the small data set we generate 100 random data splits and report average partial ROC curves and average TPR and FPR from the selected graphs. Figure 3 summarizes results for the four pairwise comparisons of GeneNet, SEM_L , GL_λ and NoShrink with ShrinkNet for the apoptosis pathway in the Ovarian cancer data set. Simulation results for the p53 pathway for the Breast cancer data are provided in SM Section 7. Table 2 and Supplementary Table 2 summarize the number of selected edges in the small and large data sets for each method.

The number of selected edges differs a lot between GeneNet, SEM_L , GL_λ and ShrinkNet (Table 2). GeneNet is the most conservative approach whereas ShrinkNet selects more edges than others in the small data sets. However, when the sample size is large GL_λ selects more than ShrinkNet, as illustrated by the number of discoveries in the large data sets. It is in-

interesting to see in Table 2 that ShrinkNet is remarkably stable in selection. The variability (as measured by the standard deviations) of the number of selected edges is relatively low, and in fact surprisingly constant in the small and large data sets, regardless of the number of selected edges. Conversely, GL_λ exhibits relatively larger variability and also large differences in number of edges.

	$n_{\text{small}}^{\text{apopt}} =$ 158	$n_{\text{large}}^{\text{apopt}} =$ 379	$n_{\text{small}}^{\text{apopt}} =$ 79	$n_{\text{large}}^{\text{apopt}} =$ 458	$n_{\text{small}}^{\text{apopt}} =$ 40	$n_{\text{large}}^{\text{apopt}} =$ 497
ShrinkNet	62.5 (5.7)	138.6 (5.9)	31.4 (5.1)	166.9 (6)	18.2 (4.8)	179.6 (5.6)
SEM _L	16.0 (3.9)	54.0 (5.3)	4.7 (2.3)	65.1 (4.7)	1.6 (1.2)	69.2 (4)
GL	25.8 (10.6)	145.7 (35.5)	9.6 (4.7)	224.1 (56)	5.3 (3.2)	282.2 (58.1)
GeneNet	10.2 (4.6)	22.9 (4.6)	2.2 (2.3)	25.8 (3.5)	0.3 (1.5)	26.1 (2.4)

TABLE 2

Average number of selected edges (and standard deviations in parentheses) for each method in the small and large data sets over 100 random partitioning of the Ovarian cancer data set.

The results in Figure 3 suggest that ShrinkNet compares very favourably to the other methods in recovering the benchmark edge set \mathcal{S}_b . In particular, edge selection (as represented by dots in the ROC plots) is shown to outperform the other methods clearly in all situations. In the most high-dimensional case $n_{\text{small}}^{\text{apopt}} = 40$, GeneNet, SEM_L and GL_λ detect almost no edges in the small data set (see Table 2), whereas ShrinkNet still detects a non-negligible number of edges, which translates into a higher TPR (with negligible FPR). Partial ROC curves in Figure 3 also indicate that edge ranking as provided by ShrinkNet is often superior to others. This is particularly true when $n_{\text{small}}^{\text{apopt}} = 79$ and $n_{\text{small}}^{\text{apopt}} = 40$. In case $n_{\text{small}}^{\text{apopt}} = 158$, SEM_L and GL_λ outperform ShrinkNet for edge ranking, but not for edge selection. This suggests that the selection procedure proposed in Section 2.4 is robust to the edge ranking on which it is based. This is confirmed by comparing ShrinkNet with NoShrink, where there is no difference in selection performance, whereas edge ranking appears to be improved by the global shrinkage prior.

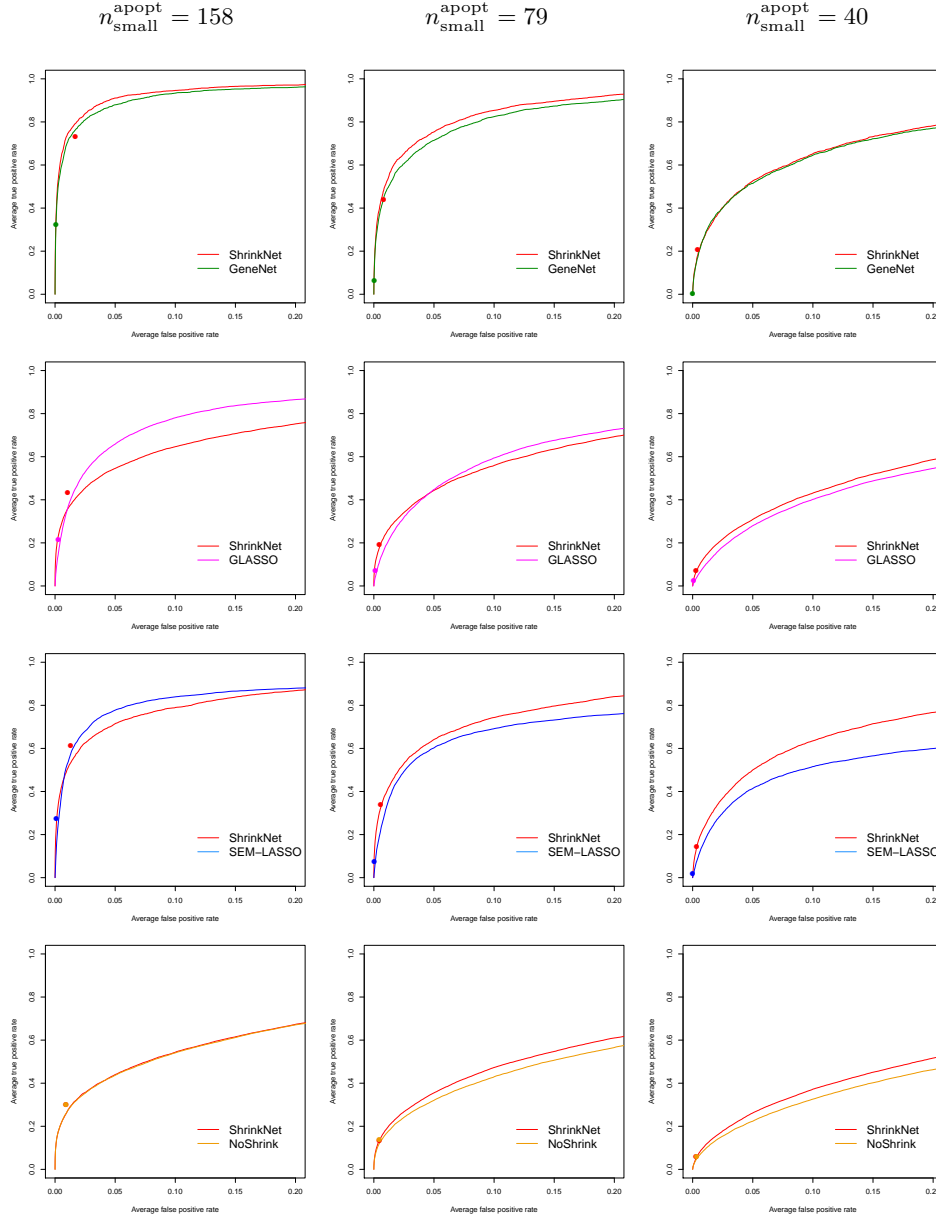


FIG 3. Average partial ROC-curves corresponding to all pairwise comparisons of GeneNet (green), GL_{λ} (magenta) and SEM_L (blue) and NoShrink (orange) with ShrinkNet (red) when the apoptosis data are randomly split into a small data set of size $n_{\text{small}}^{p53} \in \{134, 67, 34\}$ and a large validation one of size $n_{\text{large}}^{p53} \in \{392, 459, 492\}$. For each plot, dots correspond to average TPR and FPR of selected graph structures as obtained by the two inference methods under comparison.

Finally Figure 4 displays rank correlation of edges between all pairs of data sets of size $n_{\text{small}}^{\text{apopt}}$ for ShrinkNet and NoShrink. The correlations are clearly higher for ShrinkNet than for NoShrink when $n_{\text{small}}^{\text{apopt}} \in \{79, 40\}$, which indicates that the global shrinkage improves the stability and, hence, reproducibility of edge ranking when the sample size $n_{\text{small}}^{\text{apopt}}$ is not large.

In summary, the simulation study demonstrates that ShrinkNet reproduces better than SEM_L , GL_λ and GeneNet on real data.

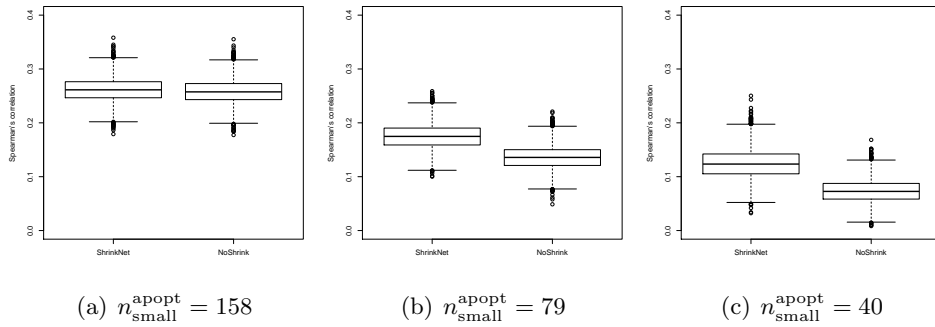


FIG 4. *Correlations of edge ranking as provided by ShrinkNet and NoShrink across the 100 generated small data sets of size $n_{\text{small}}^{\text{apopt}}$. Each boxplot displays Spearman rank correlations between the values of $\bar{\kappa}_r$, $r = 1, \dots, P$, obtained from all the $(100 \times 99)/2 = 4950$ pairs of data sets of size $n_{\text{small}}^{\text{apopt}}$ for each of the two methods. Note that one does not expect high rank correlation when considering all edges.*

4.2. Stability. In this section, the random splitting strategy is used to study the stability of edges selected by each method. Let $\hat{\pi}_{ij}$ be the empirical selection probability of edge (i, j) for a given method over the 100 generated small data sets of size $n_{\text{small}}^{\text{apopt}}$. We define the set of stable edges by $S_{\text{stable}} = \{(i, j) : \hat{\pi}_{ij} \geq \pi_{\text{thr}}\}$ where $0.5 < \pi_{\text{thr}} \leq 1$. To determine an appropriate cut-off π_{thr} , which is comparable between methods, we use the stability criterion proposed by (Meinshausen and Bühlmann, 2010). This is based on the following upper bound on the expected number $\mathbb{E}(V)$ of falsely selected edges:

$$(4.1) \quad \mathbb{E}(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)P},$$

where q is the expected number of edges selected by the given method and P is the total number of edges ($P_{\text{apopt}} = 3081$ and $P_{\text{p53}} = 2211$). To compare the set of stable edges between the different methods, we set $\mathbb{E}(V) = 30$

as in [Meinshausen and Bühlmann \(2010\)](#). Then, π_{thr} (and hence S_{stable}) is determined using an empirical estimate of q (see Table 2 and SM Table 2). Because the type I error is controlled in the same way for all methods, comparison can reasonably be based on the number of stable edges.

To illustrate, when $n_{\text{small}}^{\text{apopt}} = 158$ for the apoptosis data we obtain that $\pi_{\text{thr}}^{\text{ShrinkNet}} = 0.623$, $\pi_{\text{thr}}^{\text{SEM}_L} = 0.508$, $\pi_{\text{thr}}^{\text{GL}_\lambda} = 0.522$ and $\pi_{\text{thr}}^{\text{GeneNet}} = 0.503$, which result in 27, 12, 12 and 8 stable edges, respectively. These are illustrated in the left column of Figure 5. As $\mathbb{E}(V)$ is fixed, the value of π_{thr} only varies between methods because estimates of q differ. This is intuitive: if the method selects a lot of (few) edges we expect π_{thr} to be large (small).

Figure 5 and SM Figure 10 display stable edges obtained with each method as a function of $n_{\text{small}}^{\text{apopt}}$ and $n_{\text{small}}^{\text{p53}}$, respectively. For the two data sets ShrinkNet selects an important number of stable edges. This is particularly true for the apoptosis pathway where the method clearly yields more stable edges than SEM_L , GL_λ and GeneNet in all situations. Specifically, when $n_{\text{small}}^{\text{apopt}} = 79$ ShrinkNet identifies a nearly identical network to GL_λ and SEM_L when $n_{\text{small}}^{\text{apopt}} = 158$. For the p53 pathway (see SM Figure 10), GL_λ detects more stable edges than ShrinkNet when $n_{\text{small}}^{\text{p53}} = 134$, as many as when $n_{\text{small}}^{\text{p53}} = 67$, and less when $n_{\text{small}}^{\text{p53}} = 40$. This suggests that when the sample size is small ShrinkNet tends to select more stable edges than GL_λ . Finally, for the two data sets ShrinkNet detects more stable edges than SEM_L and GeneNet.

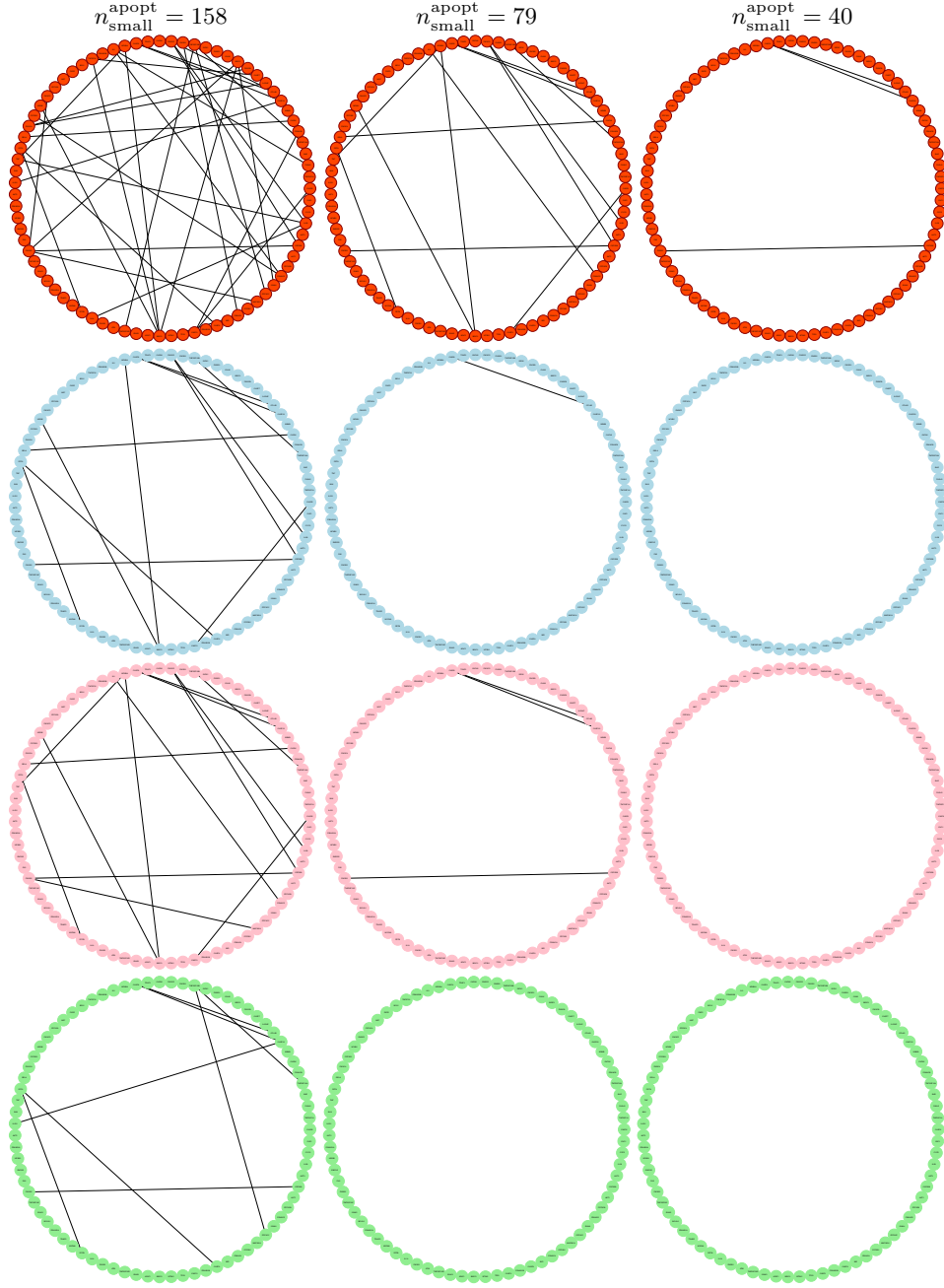


FIG 5. Stable edges for the apoptosis pathway obtained with ShrinkNet (red), SEM_L (blue), GL (pink) and GeneNet (green) when $\mathbb{E}(V) = 30$ as a function of $n_{\text{small}}^{\text{apopt}}$. Plots were generated using the R CRAN package rags2ridges.

5. Conclusion. In this paper we proposed a Bayesian SEM with global-local shrinkage priors for gene network reconstruction. The model employs simple conjugate priors to impose regularization. Because these are not sparse, a novel method for a posteriori edge selection was introduced to infer the graph structure. Computational efficiency was achieved by SVD decompositions and fast variational approximations. We discussed empirical Bayes estimation of prior hyper-parameters and embedded this in a variational EM-type algorithm. The simulations showed that the proposed approach is often superior to popular (sparse) methods in low-, moderate- and high-dimensional cases. In particular, on real data the method yielded more stable and reproducible discoveries.

A novelty of our work is the use of *global* shrinkage priors, which allow the borrowing of information across regression equations. We are not aware of any previous works combining global and local shrinkage priors. In the frequentist setting Yuan et al. (2012) borrows information across the regularizing parameters corresponding to ℓ_1 -penalties by combining local and global searches. In the Bayesian setting the focus is often on studying the equivalence between the SEM and a proper joint distribution (Geiger and Heckerman, 2002; Dobra et al., 2004). In this paper we have shown that the combined use of global and local shrinkage priors improves statistical inference, in particular edge ranking.

Our variable selection method performs simultaneous selection of the two parameters that are associated with each edge, but unlike sparsity-based methods performs separate estimation and selection steps. However, separating estimation and selection may also come as an advantage in terms of optimizing performance with respect to either of these criteria. In fact, “The idea of pre-ranking covariates and then selecting models has become a well established technique in the literature” (Ishwaran and Rao, 2005, Remark 6).

An important practical advantage of our approach is that the estimation procedure is coherent and complete, and does not rely on tuning, resampling, or cross-validation to set regularization parameter(s). This is particularly encouraging for extending the method to settings with multiple types of high-dimensional covariates, which would require different amounts of shrinkage. For methods based on resampling or cross-validation this may become overly computationally burdensome.

The proposed method is particularly suitable for gene network reconstruction using expression data. This type of network aims at providing a picture of regulatory mechanisms that act between genes. In practice, the interest often lies in a relatively small subset of genes that are known to be func-

tionally linked (e.g. a pathway). In this context the Bayesian SEM may be more appropriate than others, because such a gene set is usually of moderate dimension and, hence, due to the functional link, the corresponding network is likely to be relatively less sparse. Therefore strong dependencies between genes are more likely to occur and this may favor Normal-Gamma (ridge-type) regularization. In addition, the coherence in functionality may render shrinkage beneficial for parameter estimation in the SEM.

We have focused on recovering the support of the precision matrix, but it is also possible to obtain an estimate of it. An immediate approach is to use the graph structure provided by the Bayesian SEM as a prior for precision estimation (sometimes referred to as *parameter learning* (Scutari, 2013)). Versions of the Wishart distribution, such as the G-Wishart (Dobra, Lenkoski and Rodriguez, 2011; Wang and Li, 2012), are computationally attractive. Other estimation strategies have been proposed outside the Bayesian paradigm. These are usually based on the idea of thresholding. See, for example, Zhou et al. (2011) and Yuan (2010).

We foresee several extensions. SEMs are appropriate to describe directed networks and it would be interesting to investigate different types of shrinkage priors suitable in this context, for example to shrink in- and outgoing edges differently. Extension to non-Gaussian data is possible, where it may be desirable to adopt a flexible likelihood model and other types of posterior approximations may be considered (Rue, Martino and Chopin, 2009). Finally the model suits construction of integrative networks when allowing different priors for different types of interactions.

SUPPLEMENTARY MATERIAL

Technical details and complementary results

(DOI link here). We present technical and methodological details regarding the variational approximation and the different methods under comparison in Sections 3 and 4. Furthermore, complementary simulation results are provided.

References.

- ALLEN, G. I. and LIU, Z. (2013). A Local Poisson Graphical Model for Inferring Networks From Sequencing Data. *NanoBioscience, IEEE Transactions on* **12** 189-198.
- BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1** 121-144.
- BONDELL, H. D. and REICH, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.* **107** 1610-1624. [MR3036420](#)

- BRAUN, M. and MCAULIFFE, J. (2010). Variational Inference for Large-Scale Models of Discrete Choice. *Journal of the American Statistical Association* **105** 324–335.
- CERAMI, E., GAO, J., DOGRUSOZ, U., GROSS, B. E., SUMER, S. O., AKSOY, B. A., JACOBSEN, A., BYRNE, C. J., HEUER, M. L., LARSSON, E., ANTIPIN, Y., REVA, B., GOLDBERG, A. P., SANDER, C. and SCHULTZ, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* **2** 401–404.
- CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771.
- CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64.
- DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. [MR2896846](#)
- DOBRA, A., HANS, C., JONES, B., NEVINS, J. R., YAO, G. and WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. [MR2064941](#)
- DODD, L. E. and PEPE, M. S. (2003). Partial AUC estimation and regression. *Biometrics* **59** 614–623. [MR2004266](#)
- EFRON, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Institute of Mathematical Statistics Monographs 1. Cambridge: Cambridge University Press. xii, 263 p.
- FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems 23* (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) 604–612.
- FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GAO, X., PU, D. Q., WU, Y. and XU, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statist. Sinica* **22** 1123–1146. [MR2987486](#)
- GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. [MR1936324](#) (2003h:62019)
- GIRAUD, C. (2008). Estimation of Gaussian graphs by model selection. *Electron. J. Stat.* **2** 542–563. [MR2417393](#) (2009m:62114)
- ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773.
- JACOBSEN, A. (2013). cgdsr: R-Based API for accessing the MSKCC Cancer Genomics Data Server (CGDS). R package version 1.1.30.
- KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90** 928–934.
- KRÄMER, N., SCHAFER, J. and BOULESTEIX, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics* **10** 384.
- LEDOT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. [MR2026339](#) (2004m:62130)
- LIAN, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *J. Statist. Plann. Inference* **141** 2839–2848. [MR2787749](#) (2012c:62228)
- LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of

- g-priors for Bayesian variable selection. *Journal of the American Statistical Association* **103** 410–423.
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363 \(2008b:62044\)](#)
- MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 417–473. [MR2758523](#)
- MOHAMMADI, A. and WIT, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10** 109–138.
- OATES, C. J. and MUKHERJEE, S. (2012). Network inference and biological dynamics. *Ann. Appl. Stat.* **6** 1209–1235.
- ORMEROD, J. T. and WAND, M. P. (2010). Explaining variational approximations. *Amer. Statist.* **64** 140–153. [MR2757005](#)
- PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. [MR2524001](#)
- PENG, J., ZHOU, N. and ZHU, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.* **104** 735–746. [MR2541591 \(2011c:62061\)](#)
- RAJAGOPALAN, M. and BROEMELING, L. (1983). Bayesian inference for the variance components in general mixed linear models. *Comm. Statist. A—Theory Methods* **12** 701–723. [MR696816 \(84g:62112\)](#)
- RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Ann. Statist.* **38** 1287–1319.
- RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 319–392. [MR2649602](#)
- SCHAEFER, J., OPGEN-RHEIN, R. and STRIMMER, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News* **6/5** 50–53.
- SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28 pp. (electronic). [MR2183942](#)
- SCUTARI, M. (2013). On the prior and posterior distributions used in Graphical modelling. *Bayesian Analysis* **8** 1–28.
- VALPOLA, H. and HONKELA, A. (2006). Hyperparameter adaptation in variational Bayes for the gamma distribution Technical Report, Helsinki University of Technology.
- VAN DE WIEL, M. A., LEDAY, G. G. R., PARDO, L., RUE, H., VAN DER VAART, A. W. and VAN WIERINGEN, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14** 113–128.
- VAN WIERINGEN, W. N. and PEETERS, C. F. W. (2014). Ridge estimation of inverse covariance matrices from high-dimensional data. *ArXiv e-prints*.
- WANG, H. and LI, S. Z. (2012). Efficient Gaussian graphical model determination under G -Wishart prior distributions. *Electron. J. Stat.* **6** 168–198. [MR2879676](#)
- WARTON, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Amer. Statist. Assoc.* **103** 340–349. [MR2394637 \(2009e:62235\)](#)
- WEST, M. (2003). Bayesian factor regression models in the “large p , small n ” paradigm. In *Bayesian statistics, 7 (Tenerife, 2002)* 733–742. Oxford Univ. Press, New York. [MR2003537](#)
- YANG, E., RAVIKUMAR, P., ALLEN, G. and LIU, Z. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25* (P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger,

- eds.) 1367–1375.
- YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. [MR2719856 \(2011k:62149\)](#)
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824](#)
- YUAN, Y., CURTIS, C., CALDAS, C. and MARKOWETZ, F. (2012). A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Trans Comput Biol Bioinform* **9** 947–954.
- ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain)* (J. M. BERNARDO, M. H. DEGROOT, D. V. LINDLEY and A. F. M. SMITH, eds.) **1**. Valencia: University Press.
- ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The **huge** package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. [MR2930633](#)
- ZHOU, S., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.* **12** 2975–3026. [MR2854354](#)

MRC BIOSTATISTICS UNIT
CAMBRIDGE INSTITUTE OF PUBLIC HEALTH
FORVIE SITE
ROBINSON WAY
CAMBRIDGE BIOMEDICAL CAMPUS
CAMBRIDGE CB2 0SR
UNITED KINGDOM
E-MAIL: gwenael.leday@mrc-bsu.cam.ac.uk

DEPARTMENT OF MATHEMATICS
VRIJE UNIVERSITEIT AMSTERDAM
DE BOELELAAN 1081
1081 HV AMSTERDAM
THE NETHERLANDS
E-MAIL: degunst@cs.vu.nl
E-MAIL: w.vanwieringen@vumc.nl
E-MAIL: mark.vdwiels@vumc.nl

MATHEMATICAL INSTITUTE
FACULTY OF SCIENCE
LEIDEN UNIVERSITY
P.O. Box 9512
2300 RA LEIDEN
THE NETHERLANDS
E-MAIL: avdvaart@math.leidenuniv.nl
E-MAIL: g.b.kpogbezan@math.leidenuniv.nl

DEPARTMENT OF EPIDEMIOLOGY AND BIostatISTICS
VU UNIVERSITY MEDICAL CENTER
PO Box 7057
1007 MB AMSTERDAM
THE NETHERLANDS
E-MAIL: w.vanwieringen@vumc.nl
E-MAIL: mark.vdwiels@vumc.nl